

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-014991

(43)Date of publication of application : 18.01.2002

(51)Int.Cl.

G06F 17/30

G06F 13/00

(21)Application number : 2000-193794

(71)Applicant : HITACHI LTD

(22)Date of filing : 28.06.2000

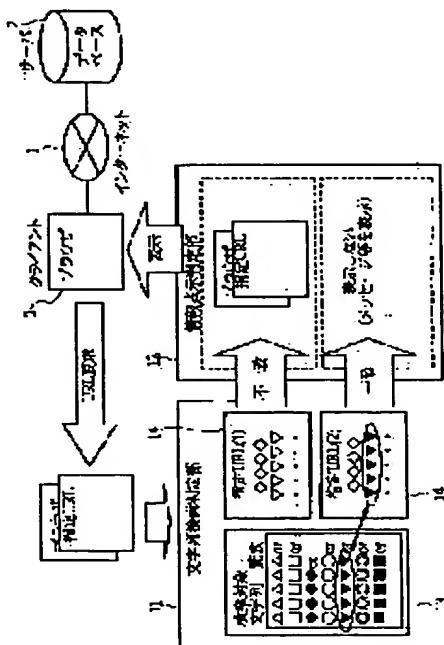
(72)Inventor : KATO SHINGO

(54) INFORMATION FILTERING DEVICE ON NETWORK

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an information filtering device on a network for extracting only the proper information by limiting accesses to the information which are improper to users in regard to a manager or internet users.

SOLUTION: A character string retrieval filtering system includes a character string retrieval decision part 11 which decides whether the retrieval condition consisting of a prescribed character string is included in each document of pages making up the retrieved information before this information is displayed on a client, an information display decision part 12 which decides whether the document should be displayed on the client in each of contents of the retrieval condition and for each information when the deciding result of the part 11 shows that the relevant retrieval condition is included in the document and other parts. In such a constitution, the retrieval/comparison is carried out for and between a retrieval object character string list 13 and the texts included in designated URL (1) and (2) 14 and the pages including the harmful information are not displayed.



LEGAL STATUS

[Date of request for examination] 05.08.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2002-14991

(P2002-14991A)

(43)公開日 平成14年1月18日 (2002.1.18)

(51)Int.Cl. ¹ G 0 6 F 17/30 13/00	識別記号 3 4 0 1 1 0 5 4 0	F I G 0 6 F 17/30 13/00	テマコト [*] (参考) 3 4 0 A 5 B 0 7 5 1 1 0 F 5 4 0 E
--	---------------------------------	-------------------------------	---

審査請求 未請求 請求項の数1 O L (全7頁)

(21)出願番号 特願2000-193794(P2000-193794)

(22)出願日 平成12年6月28日 (2000.6.28)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 加藤 審吾

神奈川県海老名市下今泉810番地 株式会

社日立製作所P C事業部内

(74)代理人 100080001

弁理士 筒井 大和

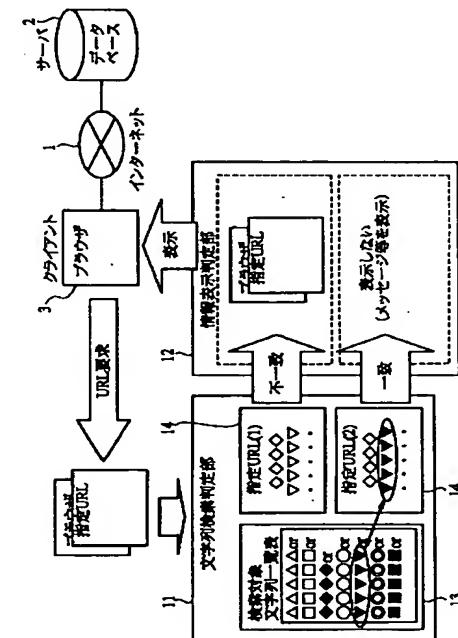
Fターム(参考) 5B075 KK63 ND03 ND36 PQ02

(54)【発明の名称】 ネットワーク上の情報フィルタリング装置

(57)【要約】

【課題】 管理者またはインターネットユーザから見て、ユーザにとって不適切な情報へのアクセスを制限して、適切な情報のみを抽出することができるネットワーク上の情報フィルタリング装置を提供する。

【解決手段】 文字列検索フィルタリングシステムであって、検索された情報をクライアント上に表示する前に、この情報を構成する各ページの文書に対して、所定の文字列からなる検索条件が含まれるか否かを判定する文字列検索判定部11と、判定の結果、検索条件が文書に含まれるときは、この検索条件の内容毎に文書を情報単位毎にクライアント上に表示するか否かを判定する情報表示判定部12などから構成され、検索対象文字列一覧表13と指定URL(1),(2)14に含まれるテキストとの検索/比較を行い、有害な情報が含まれているページを表示させないようにする。



【特許請求の範囲】

【請求項1】 ネットワーク上に存在するサーバの所定の情報をクライアントがアドレスを指定して検索し、この検索された情報をフィルタリングするネットワーク上の情報フィルタリング装置において、前記検索された情報を前記クライアント上に表示する前に、この情報を構成する各ページの文書に対して、所定の文字列からなる検索条件が含まれるか否かを判定する手段と、

前記判定の結果、前記検索条件が前記文書に含まれるときは、この検索条件の内容毎に前記文書を情報単位毎に前記クライアント上に表示するか否かを判定する手段と、を有することを特徴とするネットワーク上の情報フィルタリング装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、電子または光等を媒体とする記憶装置や情報通信網から情報を取り出す際に、不要もしくは不適切な情報へのアクセスを制限する情報フィルタリング技術に関し、たとえばインターネット上に存在するサイト(URL)検索をブラウザにて行う場合、そのブラウザ上にサイトの情報が表示される前にフィルタリングを行う技術として好適なネットワーク上の情報フィルタリング装置に適用して有効な技術に関する。

【0002】

【従来の技術】 従来、インターネット上に存在するサイト(URL)検索をブラウザにて行う場合、そのサイトの情報に含まれる有害な情報の閲覧を規制するには、管理者がそれら有害な情報を含むインターネットサイト(URL)アドレスのデータベースを作成し、もしくはそれらデータベースを提供している会社からデータベースを購入し、そのデータベースを元にサーバ側で有害な情報を制限する技術が用いられている。

【0003】 なお、このようなインターネット上に存在する有害な情報を制限する技術としては、たとえば1999年8月16日、日経BP社発行の「日経コンピュータ(no. 476)」P154~P156等の文献に記載される技術が挙げられる。

【0004】

【発明が解決しようとする課題】 しかしながら、前記のような技術では、一般家庭でブラウジングを行う際に、簡単かつ有効的に有害サイトをフィルタリングすることができない。すなわち、従来の方法でURLのフィルタリングを行うには、サーバの管理者もしくはブラウザの使用者が表示させたくない(もしくはしたくない)URLのアドレスを直接指定したデータベースを作成し(もしくはデータベースを購入し)、それら有害なURLを表示させないようにしているため、データベースに登録されていない有害なサイトを規制(フィルタリング)す

ることはできない。しかも、データベースの更新も頻繁に行わなくてはならない。

【0005】 詳細に、従来の情報フィルタリングでは、たとえばWWW上のWebページ等に適用する場合においては、以下に示すような問題が存在していた。

【0006】 (1) Webページは単一の情報からなる場合と複数の情報からなる場合があり、複数の情報からなるページの場合に、個々の情報単位毎に分割し、その情報単位毎にプロファイルとの比較を行なわないと、不必要的情報のフィルタリングが正確にできない。

【0007】 (2) 大規模なシステムでない場合、全世界のページを網羅的にチェックすることは単独システムでは不可能である。Webページはハイパーテキストであるために、複数のページによって一定の情報を表現することがあり、前述のフィルタ手段が一つのWebページだけしか指定できないと、そのページからリンクを張られている子供ページや孫ページに含まれる有害情報はフィルタリングできない。

【0008】 (3) 単独のフィルタリング機能の処理だけでは、利用者にとって十分な範囲の新規発生情報をフィルタリングすることが困難である。

【0009】 また、他方で、それらを管理するサーバ等を経由しない一般家庭では、ユーザにとって不適切な情報へのアクセスを制限することができないという課題を有していた。

【0010】 そこで、本発明は、上記のような実情に鑑みて為されたものであり、管理者またはインターネットユーザから見て、ユーザにとって不適切な情報へのアクセスを制限して、適切な情報のみを抽出することのできる文字列検索フィルタリング技術を提供することを目的とするものである。また、本発明の機能は、管理サーバ側もしくはクライアント側のブラウザのどちらにも容易に組み込むことを可能とするものである。

【0011】 詳細に、本発明は、上記のWWW上のWebページ等に適用するような背景を考慮したものであり、WWWのように個々人が独自にデータを作成および修正するデータベースにおいて、利用者にとって有害な情報のみを効率的にフィルタリングして通知しないようにすることを可能とする文字列検索フィルタリング技術を提供するものである。

【0012】

【課題を解決するための手段】 本発明は、ネットワーク上に存在するサーバの所定の情報をクライアントがアドレスを指定して検索し、この検索された情報をフィルタリングする装置に適用され、検索された情報をクライアント上に表示する前に、この情報を構成する各ページの文書に対して、所定の文字列からなる検索条件が含まれるか否かを判定する手段と、この判定の結果、検索条件が文書に含まれるときは、この検索条件の内容毎に文書を情報単位毎にクライアント上に表示するか否かを判定

する手段と、を有することを特徴とするものである。
【0013】詳細に、本発明の文字列検索フィルタリング機能は、インターネット上に存在するサイト（URL）検索をブラウザにて行う場合、そのブラウザ上にサイトの情報が表示される前にフィルタリングする機能であり、予め登録された検索条件（文字列／文字コード：検索条件は複数指定可能）と、判定される文書（インターネットサイトのハイパーテキスト等）に含まれる情報との間の類似度を算出し、その算出した類似度に従って文書の中から所定の文字列を直接フィルタリングする文字列検索フィルタリング機能において、前記文書に複数の検索条件を含むか否かを判定する手段を備えているものである。

【0014】この発明の文字列検索フィルタリング機能においては、ブラウザに情報が表示される前にフィルタリング機能が、Webページの文書それぞれに対して、検索条件である文字列からなるデータが検索対象のWebページの文書に含まれるかどうかを判定する。そして、この判定機能によって検索条件（文字列）が含まれるデータと判定されたときに、その内容毎にフィルタリング処理を行なうべく文書を情報単位毎にブラウザ上に表示するか否かを判定する。これにより、この発明の文字列検索フィルタリング機能では、単一の内容からなるWebページと複数の内容からなるWebページに対し、これら全てをフィルタリング対象とし、かつ内容に応じた高精度のフィルタリングを可能とすることができる。

【0015】また、本発明の文字列検索フィルタリング機能は、複数の文書の中から所定の文字列を選出する文字列検索フィルタリング機能であって、階層構造をなすハイパーテキストをフィルタリング対象の文書として、それらに含まれる検索対象文字列において、新たな情報が発生した場合においても本機能により登録された文字列を元に下位層に位置する文書に対するフィルタリングをすることが可能である。

【0016】これらの機能によって、設定されたアドレスに関係なく、ページ毎にフィルタリングが可能なので、URLのアドレス指定のフィルタリングに比べ、その範囲内外に新たな情報が発生した場合においてもフィルタリングすることが可能となる。

【0017】以上のように、本発明の文字列検索フィルタリング機能においては、フィルタすべき文字列を設定／指定することにより、その設定／指定された文字列を起点としてフィルタリングを行うので、階層化されているWebページもフィルタ対象とし、全てのブラウジング範囲のデータを対象にフィルタリング処理を行なう。これにより、階層的なWebページ等のフィルタリングも可能とし、指定した範囲内に新規または修正された情報がある場合にも、それらをもれなく検知／フィルタリングすることができる。

【0018】また、ブラウザよりインターネットサイト（URL）（ブラウザには表示させたくない有害なURL）への接続要求があった場合、文字列検索フィルタリング機能により、表示させたくない文字列等がURLの指示示すページ上に存在した場合は、その指示示すページを表示させないようにすることができる。

【0019】さらに、文字列検索によるURLフィルタリング機能は、URLのアドレスを直接指定しなくとも、そのURLが指示示すテキスト内に含まれる文字列により、表示させたくない有害な情報等を含むURLをフィルタリングすることができる。

【0020】

【発明の実施の形態】以下、本発明の実施の形態を図面に基づいて詳細に説明する。図1は本発明の一実施の形態の文字列検索フィルタリングシステムを示す概略構成図、図2は本実施の形態の文字列検索フィルタリングシステムにおいて、文字列検索フィルタリング処理の流れを示すフロー図である。

【0021】まず、図1により、本実施の形態の文字列検索フィルタリングシステムの一例の構成を説明する。

本実施の形態の文字列検索フィルタリングシステムは、たとえばインターネット1上に接続されたサーバ側のデータベース（サイト）2と、クライアント側のブラウザ3などからなる構成において、クライアント側に構築され、検索された情報をクライアント上に表示する前に、この情報を構成する各ページの文書に対して、所定の文字列からなる検索条件が含まれるか否かを判定する文字列検索判定部11と、判定の結果、検索条件が文書に含まれるときは、この検索条件の内容毎に文書を情報単位毎にクライアント上に表示するか否かを判定する情報表示判定部12などから構成されている。

【0022】詳細に、この文字列検索フィルタリングシステムにおいては、クライアント側のブラウザ3よりインターネット1上のサイト（URL）の要求があった場合に、その要求されたハイパーテキストやその他のデータを含むURLをブラウザ3に表示する前に、まず文字列検索判定部11で予め登録された検索対象となる文字列がそのURLに含まれるかどうかを検索し、検索条件が一致した場合はクライアント側のブラウザ3にはその情報を含むURLを表示せず、反対に検索条件が一致しない場合はフィルタリング対象ではないと判断して、クライアントから要求のあったURLの情報をブラウザ3に表示する処理を情報表示判定部12で行なう構成となっている。

【0023】また、検索方法に関しては、たとえば特開平11-353329号公報「文書検索方法及びその実施装置並びにその処理プログラムを記録した媒体」等の検索方法を流用でき、あらゆる検索方法を使用できるものとする。

【0024】なお、この発明は、クライアントのブラウ

ザ3の機能としても、企業や大学等のプロキシーサーバ等のサーバ機能の一部としての実施も可能であり、媒体であるフロッピディスクやCD-ROM等に格納した形態や、磁気ディスク等に格納しておいて、ネットワークで入手可能な形態で提供することも可能である。

【0025】図1を用いて、さらに本実施の形態の文字列検索フィルタリングシステムの機能を説明する。図1に示すように、本実施の形態の文字列検索フィルタリングシステムは、ユーザが任意に登録した検索対象の文字列をテキストベースにて保存した検索対象文字列一覧表13と、クライアントのブラウザ3から要求のあった指定URL(1)、(2)14に含まれるハイパーテキスト内のテキストと検索/比較する。なお、ハイパーテキスト内の全ての構成部分が検索対象となる。図1では、検索対象文字列一覧表13と指定URL(2)14に含まれるテキストで一致した文字列が見つかっている。

【0026】この検索対象文字列一覧表13は、システムが監視すべき文字列の一覧である。利用者がこの検索対象文字列一覧表13に監視したい文字列を登録する。なお、文字列とは、全ての文字コード(ASCII、ソフトJIS、JIS、EUC、他)を含むものとする。

【0027】次に、本実施の形態の作用について、図2により、文字列検索フィルタリング処理の流れを説明する。

【0028】この文字列検索フィルタリング処理は、クライアント側からURLの要求があった場合に(ステップS1)、このURLの検索を行い(ステップS2)、URLを見つけた後に(ステップS4)、インターネットのデータベース2からダウンロードされた全てのページ(ハイパーテキスト)に対して処理を行なう。なお、ステップS1において、URLの要求がない場合はなにもせず(ステップS2)、またステップS4でURLが見つからなかった場合はステップS1の処理に戻る。

【0029】まず始めに、文字列検索フィルタリングシステムは、ダウンロードされたWebページのハイパーテキストを取り出し、その取り出されたハイパーテキストをユーザが任意に登録した検索対象の文字列をテキストベースにて保存した検索対象文字列一覧表13に基づいて、登録されている検索対象文字列をOR検索条件をもとにハイパーテキスト内に含まれる文字列の検索を実行し(ステップS5)、そのページに検索条件が見つかるか否かを文字列検索判定部11で判定する(ステップS6)。

【0030】そして、ステップS6の判定の結果、検索対象文字列がページ内に含まれた場合には、情報表示判定部12で有害な情報を含まれていると判断して、対象とするページを表示させない処理を行う(ステップS7)。この非表示処理を実施した後に、処理対象のページを表示できない趣旨のメッセージをブラウザ3に表示

する(ステップS8)。

【0031】反対に、ステップS6の判定の結果、検索対象文字列がページ内に含まれない場合は、情報表示判定部12で有害な情報を含まれていないと判断して、URLに対応する処理対象のページをブラウザ3に表示する(ステップS9)。

【0032】続いて、URLの要求があったか否かを判定し(ステップS10)、URLの要求があった場合はステップS3からの処理を繰り返し、またURLの要求がない場合は終了となる。

【0033】この際に、複数の情報単位からなっているページも、本文字列検索フィルタリングシステムでは目的ページをブラウザ3に表示する前に文字列検索フィルタリング処理を行うので、サブディレクトリを含む、もしくはリンク先を含むページでも、それらのページをブラウザ3に表示する前に文字列検索フィルタリング処理を行うことが可能なので、利用者に提示する結果を高い精度でフィルタリングすることができる。

【0034】また、本実施の形態では、今回のフィルタリング時に取り込んだページと、前回のフィルタリング時に取り込んだページとを比較する必要がない。そのページに修正が施されたか否かを判定する必要もなく、変化があった場合でも、変化がなかった場合でも取り込んだページをフィルタリングする。なお、一度取り込んだページに検索対象文字列が含まれていた場合、そのページのアドレスを記録し、2度目にそのページを参照した場合は、そのアドレスをフィルタリング対象として判定を行い、処理の高速化に用いても良いことはいうまでもない。

【0035】次に、具体的なWebページの情報判定処理について説明する。ハイパーテキスト内は、一般的に、開始タグと終了タグとによって論理的な構造をしている。たとえば、HTMLでは、開始タグ<TITLE>と終了タグ</TITLE>とに囲まれた部分がタイトル、開始タグと終了タグとに囲まれた部分が箇条書きと定義されている。また、段落を規定する<P>や、箇条書きの各項目を表現するのように、終了タグを省略してよいタグも存在する。これらのタグについては、同じ開始タグが出現した時点で終了タグが存在したものと見なされる。文字列検索フィルタリングシステムでは、これらタグを指定してタグ内に含まれる情報のみを検索対象とすることも可能だが、タグ等を指定せずにHTML文に含まれる情報を全て検索対象とできる。

【0036】このようにタグを指定する場合は、検索速度を早くすることが目的である。この場合、先にページ内をスキャンしてHTMLの開始タグを検出す。そして、その開始タグに対応する終了タグを検出することにより、各タグに対応する情報を取り出し、タグ内のみを検索対象とする。

50

【0037】このような文字列検索フィルタリングシステムは、処理対象とするページが複数の情報単位からなるページであるかどうかを判断する必要がなく、ページ単位で判定することが可能である。

【0038】さらに、文字列検索の処理は、検索対象文字列一覧に格納された検索条件と処理対象となる各情報単位とをそれぞれ単語頻度のベクトルとして表現し、これらベクトル間の内積を取ることによって、類似度を求めるといった従前の算出方法を流用することも可能である。

【0039】本実施の形態では、市場で使われている一般的なHTMLブラウザで表示することも想定しているため、HTML形式で結果を出力している。これは、フィルタリング結果で選択された文書のオリジナルをアクセスする場合に、その文書形式との統一性を図るためにある。したがって、必ずしもこれに限定するものではなく、特殊なブラウザで取り込める形式のデータに変換するように作成することは、ごく簡単である。また、サーバ側としての機能にも容易に採用／組み込めるため、特殊な専用のブラウザを用意する必要はない。同様に、クライアント側のブラウザに本機能を追加／組み込んだり、また専用のブラウザを作成することも容易である。

【0040】このように、本実施の形態の文字列検索フィルタリングシステムによれば、単一の内容からなるWebページと、複数の内容からなるWebページに関係なく、これらを全てフィルタリング対象とし、かつ内容に応じた高精度のフィルタリング処理を実施することができる。

【0041】本実施の形態を用いると、設定したページの下位層に位置するページに新規情報を含むかどうかを再帰的にチェックする必要がない。

【0042】また、階層構造をなすページの最初のページに検索対象（フィルタリング対象）となる文字列が含まれていた場合、それ以下のページをたどらないようにすることも可能である。また、その下位に位置するページ毎にフィルタリングを行うことも可能である。

【0043】以上のように、本実施の形態は、小規模及び大規模など、どのようなシステムでも容易に導入することが可能である。システムに検索／監視させる文字列（文字コード）を、検索対象文字列一覧表13のリストに利用者自らが登録するので、インターネット上に存在する膨大な量のアドレスを登録する必要はない。特に、大規模なシステムである場合、監視するページの全てのアドレスを事前に登録することは困難である。また同様に、小規模のシステムの場合でも、インターネット上に存在する全ての有害な情報を含むページを事前に登録することは不可能である。そこで、取り込んだページに記述されている文字列をフィルタリングの対象とする本実施の形態である文字列検索フィルタリングシステムが有効になる。大規模システムとして実施する場合

は、この形態によって規制の範囲を拡大することも可能である。さらに、Webページでは、外部のページへリンクを張っている場合があるが、このような外部へのリンクについては無視するように変形することも可能である。

【0044】このように、本実施の形態の文字列検索フィルタリングシステムによれば、階層的に配置されたWebページも簡単にフィルタリングすることを可能とし、指定した範囲内に新規または修正された情報がある

10 場合でも、それらをもれなく検知し、フィルタリングすることが可能である。

【0045】また、本実施の形態では、処理の性能を高めるため、他の情報フィルタリング装置が出力するフィルタリング結果のファイルを、直接、本発明の機能とリンクするように変形することは容易である。

【0046】このように、本実施の形態の文字列検索フィルタリングシステムを使用すれば、他の情報フィルタリング装置が出力したフィルタリング結果を読み込むことにより、単独の文字列検索フィルタリングシステムが20 フィルタできる以上の範囲の情報をフィルタすることも可能となる。

【0047】

【発明の効果】以上詳述したように、本発明のネットワーク上の情報フィルタリング装置によれば、以下のような効果を得ることが可能となる。

【0048】(1) 複数の形態を有するWebページを始めとする文書情報のフィルタリングを統一的に処理し、利用者にとって使い易い形態で提供することができる。

30 【0049】(2) 複数の情報単位からなる文書内のフィルタリングについても、回りのテキストに影響されることなく独立して類似度を算出することもできるため、高い精度でフィルタリング処理を行なうことが可能となる。

【0050】(3) ハイパーテキスト形式の文書をフィルタリング対象とすることにより、複数のWebページで一つの情報を表現しているWebページでも効果的にフィルタリングさせることができ、また無制限に階層をたどることを排除することができるため、処理時間を抑えることも可能となる。

40 【0051】(4) 文字列検索フィルタリング機能においては、単一の内容からなるWebページと複数の内容からなるWebページに対し、これら全てをフィルタリング対象とし、かつ内容に応じた高精度のフィルタリングを実現することが可能となる。

【0052】(5) 文字列（文字コード）を元にフィルタを行うので、インターネットサイト（URL）のアドレスを直接指定したデータベースを作成する必要はなく、容易にURLのフィルタリングを実現でき、また従来から行われていたようにURLのアドレスによるフィ

50

ルタリングを必要としないので、一般ユーザが不適切な情報へアクセスすることを管理者もしくはインターネット利用者の判断で制限できるようにすることが可能となる。

【0053】(6) 本文字列検索フィルタリング機能を使ったブラウザを使用すれば、一般家庭でも文字列指定による簡単なインターネットサイト(URL)のフィルタリングを実現することが可能となる。

【0054】(7) 本発明の文字列検索フィルタリング機能は、管理サーバ側(大規模システム)もしくはクライアント側のブラウザ(小規模システム)のどちらにも*

* 容易に組み込むことが可能となる。

【図面の簡単な説明】

【図1】本発明の一実施の形態の文字列検索フィルタリングシステムを示す概略構成図である。

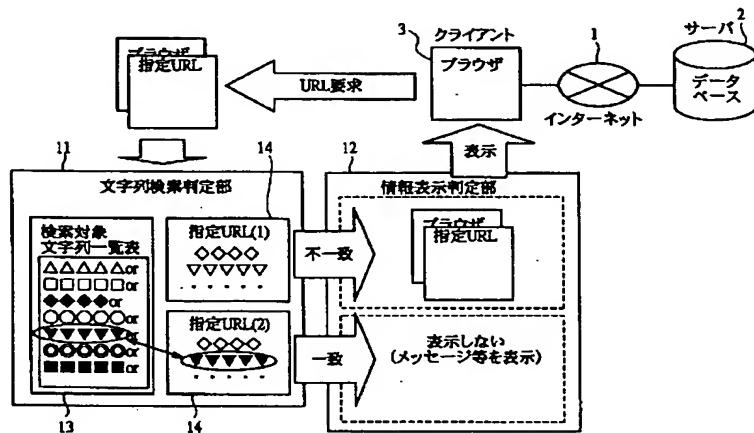
【図2】本発明の一実施の形態の文字列検索フィルタリングシステムにおいて、文字列検索フィルタリング処理の流れを示すフロー図である。

【符号の説明】

1…インターネット、2…データベース、3…ブラウザ、11…文字列検索判定部、12…情報表示判定部、13…検索対象文字列一覧表、14…指定URL。

【図1】

図1



【図2】

図2

